Title

Item statistics derived from three-option versions of multiple-choice questions are usually as robust as four- or five-option versions: implications for exam design.

Journal

Advances in physiology education, 42(4)

ISSN

1043-4046

Authors

Loudon, Catherine
Macias-Muñoz, Aide

Publication Date

2018-12-01

DOI

10.1152/advan.00186.2016

Peer reviewed

Adv Physiol Educ 42: 565–575, 2018;
doi:10.1152/advan.00186.2016.

HOW WE TEACH | *Generalizable Education Research*

Item statistics derived from three-option versions of multiple-choice questions are usually as robust as four- or five-option versions: implications for exam design

Catherine Loudon and Aide Macias-Muñoz
Department of Ecology and Evolutionary Biology, University of California-Irvine, Irvine California

Submitted 28 November 2016; accepted in final form 1 August 2018

Loudon C, Macias-Muñoz A. Item statistics derived from three-option versions of multiple-choice questions are usually as robust as four- or five-option versions: implications for exam design. *Adv Physiol Educ* 42: 565–575, 2018; doi:10.1152/advan.00186.2016.— Different versions of multiple-choice exams were administered to an undergraduate class in human physiology as part of normal testing in the classroom. The goal was to evaluate whether the number of options (possible answers) per question influenced the effectiveness of this assessment. Three exams (each with three versions) were given to each of two sections during an academic quarter. All versions were equally long, with 30 questions: 10 questions with 3 options, 10 questions with 4, and 10 questions with 5 (always one correct answer plus distractors). Each question appeared in all three versions of an exam, with a different number of options in each version (three, four, or five). Discrimination (point biserial and upper-lower discrimination indexes) and difficulty were evaluated for each question. There was a small increase in difficulty (a lower average score on a question) when more options were provided. The upper-lower discrimination index indicated a small improvement in assessment of student learning with more options, although the point biserial did not. The total length of a question (number of words) was associated with a small increase in discrimination and difficulty, independent of the number of options. Quantitative questions were more likely to show an increase in discrimination with more options than nonquantitative questions, but this effect was very small. Therefore, for these testing conditions, there appears to be little advantage in providing more than three options per multiple-choice question, and there are disadvantages, such as needing more time for an exam.

assessment; multiple choice; undergraduate

INTRODUCTION

Multiple-choice questions are one of the most common formats used in assessment (10, 13). Testing using multiple-choice questions is particularly common in large university classes, in part because of the greater ease and speed of grading of multiple-choice questions compared with other testing formats (8). In contrast to the ease of grading, it is time-consuming and difficult to compose good multiple-choice questions, especially ones that assess higher-order cognitive skills, such as critical thinking and problem-solving (8, 25). Generating meaningful alternatives to the correct answer for each question is part of the challenge of writing multiple-choice questions, and, therefore, it is helpful to know when a sufficient number of options has been reached. An excessive or unnecessarily large number of options used in a classroom exam is not only a waste of the instructor's time to produce, but also lengthens an exam for the students in an unproductive way.

The general practice in multiple-choice testing, regardless of subject matter or education level, is to provide four or five options per multiple-choice question: one correct answer and three or four "distractors," respectively (10). Undergraduate physiology textbooks often follow this convention as well, with four or five options provided in the multiple-choice questions found at the end of chapters or in the accompanying test banks. This convention gives the impression that four or five options is an optimal or desirable number from a pedagogical perspective. Owen and Froman (20) describe this rationale (with which they disagree) as "write items with 4 or 5 alternatives so that a poor student will have less chance of guessing the right answer." In fact, there is both theoretical and empirical evidence that, in some cases, three options per question (2 "distractors" plus the correct answer) are sufficient, or even preferred to a larger number of options (6, 7, 14–18, 20–24, 27–29).

In an early and influential theoretical treatment (29), it was shown that the number of possible distinct response patterns is maximized when three options are provided for each question. In that analysis, it was assumed that the total number of answer options in the entire exam was kept constant (the number of questions $\times$ the number of options/question = constant; for example a total of 120 options could be arranged in an exam as 30 questions each with 4 options, 40 questions each with 3 options, and other possibilities). This was a rough proxy for keeping the length of the exam constant. Making the reasonable assumption that the number of possible distinct response patterns relates directly to the discrimination capacity of an exam, this suggests that the discrimination capacity of an exam will be maximized when three options are provided for each question (29). That is, three options are not only enough, but actually superior to a larger or smaller number of options, when the changing number of questions is also taken into account. A related point is that the probability of attaining a perfect score on an exam by randomly guessing is minimized with three options (while holding the total number of options in the entire exam constant) (29). While these are interesting results, an instructor may not find this sufficiently compelling

Address for reprint requests and other correspondence: C. Loudon, Dept. of Ecology and Evolutionary Biology, 321 Steinhaus Hall, University of California-Irvine, Irvine, CA 92697-2525 (e-mail: cloudon@uci.edu).

to dictate the testing format in an upper-level undergraduate physiology class for biology majors because of the implicit emphasis on stochastic issues such as guessing. In addition, Grier (11) pointed out more recently that a prediction of three as the optimal number of options is only true for the case in which zero time is spent reading a question and all of the time is spent reading the options. The predicted optimal number of options per question actually increases as more time is required to read the question (relative to the options) (11). Thus there is not a fixed theoretical optimal number of options per multiple-choice question that is independent of question and option content.

There is also mixed empirical support for using three options per question. A review of responses on three different types of standardized exams (ACT, medical education program for physicians, and state certification examination in the health sciences) suggested that exam items seldom contained more than three "useful" options, meaning that the additional distractors do not function well (14). Well-functioning distractors may be identified from the pattern of distractor selection: a well-functioning distractor is one that shows a monotonically decreasing representation with score group, such that students who score higher overall are less likely to select that particular distractor (14). Rodriguez (21) summarized over 80 yr of analyses comparing three-, four-, and five-option multiple-choice questions, and concluded that three-option items are preferred because they are just as effective in testing as four- or five-option questions, but they take less time. Therefore, more questions may be used, strengthening the exam as a whole. Haladyna and Downing (15) reviewed 96 theoretical and empirical studies on multiple-choice exams and concluded that three-option multiple-choice questions suffice in most circumstances. However, much of the empirical research reviewed in those papers was for K–12 classes, had small sample sizes, involved comparisons between years, were for exams that did not count for a grade, or were for students in nonscientific subjects, and so it was unclear if that empirical evidence would be completely relevant for testing in physiology classes at the university level. Undergraduate students in physiology classes at the university level are often highly motivated students who have already passed a number of competitive prerequisite classes. Furthermore, Case and collaborators (2, 3) provide evidence that medical physiology questions that provide an extended list of options (9–23 options per question) are more effective than five options for questions that describe a patient's symptoms followed by a lengthy list of possible diagnoses (the options). Hence, this is a counterexample that suggests that, for some kinds of questions, more than three options may be preferable.

In the absence of a clear and compelling consensus, the number of options provided for multiple-choice questions was evaluated in an undergraduate class in human physiology. To evaluate whether the effectiveness of multiple-choice questions was different with three, four, or five options provided per question, three different versions of exams were prepared and given in the classroom as part of normal testing. Each version of the exam had 30 questions, in which 10 had 3 options (2 distractors), 10 had 4 options (3 distractors), and 10 had 5 options (4 distractors) (Fig. 1). The point biserial correlational coefficient and the upper-lower discrimination index were used to evaluate the discrimination ability of each question in its three different forms. Different types of multiple-choice questions may show different patterns; we tested whether quantitative questions, longer questions (more words), or questions with proportionately longer stems were more effective when a larger number of options were available.

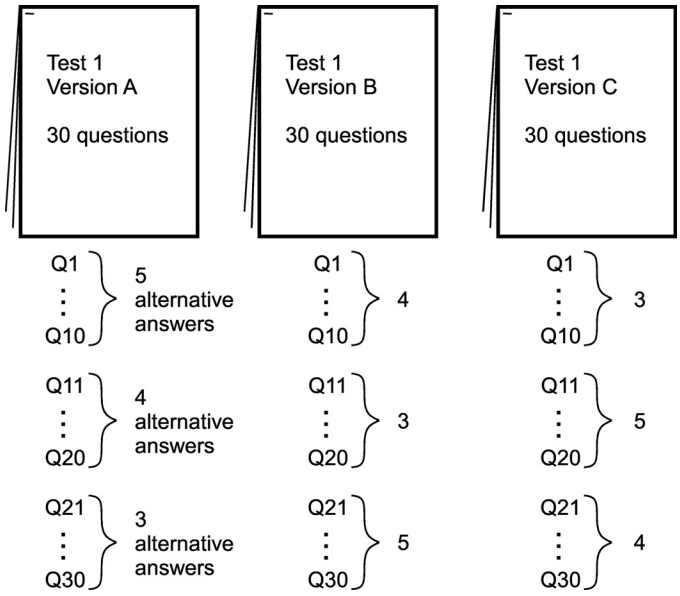

Fig. 1. Each exam came in three different versions: *A*, *B*, and *C*. One-third of the students got each version. Each version had the same 30 questions (Q1–Q30), but the number of options for a specific question varied with the version. Each exam had 10 questions with 5 options, 10 questions with 4 options, and 10 questions with 3 options, to keep all exams the same overall length. The order of the options within questions and the order of the questions were also scrambled between versions, but the unscrambled versions are shown diagrammatically for greater clarity of the experimental design.

METHODS

Testing conditions and study participants. Testing was done as part of the normal testing in the classroom. The class was Human Physiology, a lecture format course that met for 3 h/wk and lasted for a 10-wk academic quarter at a large public research university in California. The class had two sections: one section had 332 students, and the other had 330 students. The two sections had their exams on the same day during consecutive hours, and students were not allowed to keep their exams (to minimize the sharing of exam question content between students in the different sections). Students were primarily seniors (81%) and juniors (18%). The most common major of the students was Biological Sciences (70%), and the next most common major was Public Health Science (19%).

Testing protocol. All testing protocols were performed in accordance with the Family Educational Rights and Privacy Act and had Institutional Review Board approval (HS no. 2012–9201). All students had the opportunity to opt out of participation (of having their exam scores included in the analysis), none of the analyses were initiated until the final grades had been submitted, and student identifiers were removed from the data files during analysis. Participation did not affect a student's grade or score in the class. None of the students chose to opt out of the study.

Exam content and structure. There were three different versions of each exam (*versions A*, *B*, and *C*; Fig. 1) given to a section. The three versions of each exam had the same 30 questions, but differed in which questions had three, four, or five choices. On each exam version, 10 questions had 3 choices (1 correct answer and 2 distractors), 10 questions had 4 choices (1 correct answer and 3 distractors),

and 10 questions had 5 choices (1 correct answer and 4 distractors). Therefore, each version was the same total length (30 questions with a total of 120 options). The three- and four-choice versions of each question were generated from the five-choice version of the same question by discarding two or one distractor, respectively. Distractors were chosen for elimination arbitrarily (i.e., in the absence of knowing which distractors would be preferred), which had the advantage that a range was generated in the preference of eliminated distractors. Therefore, each question was answered by one-third of the students in a five-choice form, one-third of the students in a four-choice form, and one-third of the students in a three-choice form. Exam versions were randomly distributed to students: students had preassigned seats in the auditorium (randomly generated), and the stacks of exams with three alternating versions were passed out at the beginning of the class period such that students were always seated between individuals with the other versions. In addition, the order of the questions and the order of the possible answers within each question were scrambled between versions (standard procedure to decrease opportunities for cheating among students) (scrambling not shown in Fig. 1). The questions used on the exams were from a variety of sources: made up by the instructor based on questions during office hours, news items, scientific papers, or modified from different textbook or internet resources. At least one-half of the questions on each exam had not been used in this class previously, or at least for several years.

During the quarter, three multiple-choice exams were given to each of the two sections (approximately every 3 wk during the quarter), and all of the exams followed the scheme shown in Fig. 1. Therefore, there was a total of 180 questions on the 3 exams (3 exams $\times$ 2 sections $\times$ 30 questions/exam). Out of these 180 questions, there were 13 questions that were not included in this analysis because of a copying error in exam preparation, or because a second answer was deemed acceptable during the grading process. There were 29 questions that were purposefully duplicated between sections, to allow comparison between the two sections (composed of different students). Therefore, there were 138 distinct questions used in this analysis (of which 29 questions were answered by both sections).

None of these exams was cumulative: each exam covered approximately one-third of the material for the quarter. For the first two exams, students had 50 min to complete 30 multiple-choice questions (the length of the class period). The third exam had the same format and number of questions, but was given during the final exam period, and, therefore, the students had longer (1 h and 50 min) to complete that exam.

Question categorization and word counts. Questions were categorized as quantitative if they involved any calculation or comparison between numbers (e.g., figuring out which solutions would be hypotonic but hyperosmotic given information about ion concentrations and membrane permeability, or predicting how cardiac output would increase or decrease), and otherwise were categorized as nonquantitative.

Word count was used as a proxy for length of questions. The total number of words in the stem and in each option was counted for all 138 questions. The proportion of the question that was in the options was calculated as the ratio of the summed number of words in the options divided by the total number of words in that question [options/(stem + options)].

Statistical analyses. Statistical analyses were performed using SAS 9.4 (SAS Institute, Cary, NC). Mixed-model analyses were used to evaluate whether the indexes (difficulty or discrimination) were affected by the number of options (3, 4, or 5) provided on a question. There were multiple clustering factors in the data that were incorporated into the mixed model: there were two sections of the class, and, on each of the three exams, there were three versions of each exam. This means that there were 18 different student groups that answered questions in this study (2 class sections $\times$ 3 exams $\times$ 3 exam versions). The SAS procedure, Proc Mixed, was used for the mixed-model analyses using two random effects: student group (coded 1–18) and unique question identity (coded 1–138). Student group, unique question identity, and the number of options were each treated as classification variables in the mixed-model analyses. As an additional parallel analysis, two-way ANOVA of the effects of the number of options (3, 4, or 5) and the specific question (the code for one of the 138 questions) on the indexes (difficulty or discrimination) were conducted using Proc GLM, again treating both factors (the number of options and the specific question identity) as categorical variables. In all cases, the results using a mixed model with random effects were virtually identical to the general linear model without random effects, and so only the mixed-model results are reported. This pattern suggests that the clustering in the data structure was not affecting the results of the statistical analyses. Type III results are reported for all analyses (in this model, the order of the parameters does not matter, because each effect is adjusted for all other effects). For all of these analyses of the indexes (difficulty or discrimination), if a question was repeated in both class sections, only the data from a single section was used, unless noted otherwise.

T-tests were conducted in SAS using Proc TTEST. If the variances of the two samples were significantly different (at the 0.05 level), the Satterthwaite method was used to test for significant differences between the two samples. The χ^2 tests were conducted using Proc FREQ and Fisher's exact test. Analysis of covariance was used to test the effects of total word count or proportion of words that appear in the options (both treated as covariates) and number of options (a categorical variable) on difficulty or discrimination, using Proc Mixed, with both student group and question identity as random effects, as described above. Parallel analyses were performed using Proc GLM (without random effects). An $\alpha-$value of 0.05 was used to reject null hypotheses, but the actual *P* values are provided for all tests. The word "significant" is used throughout the paper to mean "statistically significant."

Calculation of indexes. Point biserial correlation coefficients (a discrimination index) were calculated for each question using the %BISERIAL macro provided by the SAS Institute (downloaded from http://support.sas.com/kb/24/991.html). The formula for the point biserial correlation coefficient for *question i* (r_i) is

$$r_i = \frac{M_{pi} - M_{qi}}{s}\sqrt{p_i q_i} \quad (1)$$

where M_{pi} is the average total exam score for the students who answered *question i* correctly, M_{qi} is the average total exam score for the students who answered *question i* incorrectly, s is the standard deviation of the total exam scores, p_i is the proportion of students who answered *question i* correctly, and q_i is the proportion of students who answered *question i* incorrectly. The maximum possible range for r_i is -1 to $+1$.

An additional discrimination index, D_i, often called the "upper-lower" index (10), was also calculated for each question. This index compares the performance of the top 27% of the students and the bottom 27% of the students for *question i*. The formula is

$$D_i = \%\text{top}_i - \%\text{bottom}_i \quad (2)$$

where $\%\text{top}_i$ is the percentage of students in the top 27% (based on the exam scores as a whole) who answered *question i* correctly, and $\%\text{bottom}_i$ is the percentage of students in the bottom 27% who answered *question i* correctly. In practice, there may be slightly more than 27%, depending on where the 27% cutoff lies (in the overall scores). Ebel and Frisbie (10) state that 27% is not significantly better than 25% or 33%, whereas Brennan (1) points out that symmetric cutoffs in the upper and lower groups are not necessary. For our questions, the average size of the upper and lower groups was 30%. The maximum possible range for D_i is -1 to $+1$.

The Kuder-Richardson 20 (KR20) was used as a statistic as one measure of a reliability estimate, usually interpreted as indicating the extent to which students taking the exam again would have the same

scores. A value of at least 0.70 is usually considered desirable. KR20 was calculated for each exam as

$$\mathrm{KR20} = \frac{N}{N-1}\left(1 - \frac{\sum p_i q_i}{s^2}\right) \quad (3)$$

where N is the number of exam items, p_i is the proportion of students answering *item i* correctly, q_i is the proportion of students answering *item i* incorrectly, and s^2 is the variance of the total exam scores. The maximum possible range for KR20 is 0 to +1. KR20 was used instead of KR21 because we had all of the information available to calculate the KR20 and did not need to use the KR21 estimate [which tends to underestimate the KR20 if the items vary in difficulty (10), which these did].

The difficulty index for a question was calculated in the standard way, as the percentage of students who answered the questions correctly (10), which means that an easier question will have a higher difficulty index. The maximum possible range for the difficulty index is 0 to +1.

Elimination of distractors might have a different outcome, depending on whether an eliminated distractor was preferred when present as an option. Because the preference could be weak or strong, we generated a new metric for quantifying the preference, rather than just classifying dichotomously (preferred vs. nonpreferred). In the five-option version of the questions, there were two distractors that were eliminated in the three-option version, and two distractors that were not eliminated in the three-option version. To quantify the preference of noneliminated vs. eliminated distractors for each question, the number of students who chose either of the noneliminated distractors minus the number of students who chose either of the eliminated distractors, all in the five-option version, was calculated for each question:

$$\begin{aligned}&\text{preference for noneliminated distractors} =\\ &\#\text{students who chose either noneliminated distractor} -\\ &\#\text{students who chose either eliminated distractor}\end{aligned} \quad (4)$$

RESULTS

The difficulty index, the percentage of the students who answered the question correctly (10), was lower when more options were given for a question ($P < 0.0001$) (Table 1). Because questions varied greatly in the magnitude of the difficulty index, question identity was included as a random effect, along with student group, in the mixed model (question identity $P < 0.0001$, student group $P = 0.06$). The magnitude of the difference in the difficulty index with more options was fairly small; 84% of the three-option questions were answered correctly, whereas 80% of the same questions with five options were answered correctly by other students in the same class section (averaged over the 138 unique questions). This small difference in difficulty index influenced by the number of options suggests that guessing does play a small role in option selection, because fewer students chose the correct answer when there were more options.

The point biserial discrimination index was not significantly affected by the number of options provided for a question ($P = 0.18$), although questions differed significantly from each other in this index ($P < 0.0001$) (Table 1). The overall average point biserial discrimination index for a question was 0.32, with a range from 0.04 to 0.53 ($n = 138$ questions; the discrimination index has a total possible range of −1 to +1). A magnitude of +0.15 is typically considered a lower limit for acceptability for this discrimination index (point biserial) (14). Eleven of the questions had a discrimination index <0.15, but most of these had an extremely high difficulty index (Fig. 2), meaning that very few students answered these 11 questions incorrectly. The point biserial discrimination index is less meaningful for a question like that (and $M_{pi} - M_{qi}$ is a very small number, see *Eq. 1*).

In contrast to the point biserial discrimination index, the upper-lower discrimination index was significantly affected by the number of options ($P = 0.02$), in addition to questions differing significantly from each other in this index ($P < 0.0001$) (Table 1). The average upper-lower discrimination index was slightly higher for the five-option version of the questions.

The relationship between the number of options, the difficulty index, and the discrimination indexes is easier to visualize by examining the response pattern for a representative question (Fig. 3). While a smaller proportion of the students who had this question in its five-option version answered it

Table 1. *Difficulty and discrimination indexes for questions comparing different numbers of options per question*

	Difficulty Index	Discrimination Index (Point Biserial)	Discrimination Index (Upper-Lower)
All data ($n = 138$ questions)			
3-option	0.84 (0.14)	0.31 (0.13)	0.25 (0.16)
4-option	0.82 (0.14)	0.31 (0.14)	0.25 (0.16)
5-option	0.80 (0.15)	0.33 (0.13)	0.27 (0.17)
Test: means are same	$P < 0.0001$*	$P = 0.18$	$P = 0.02$*
Questions parsed by difficulty index, one-half of questions (difficulty index > 0.872, $n = 69$ questions)			
3-option	0.94 (0.04)	0.27 (0.13)	0.13 (0.08)
4-option	0.93 (0.04)	0.27 (0.16)	0.15 (0.11)
5-option	0.92 (0.05)	0.30 (0.14)	0.17 (0.11)
Test: means are same	$P = 0.0002$*	$P = 0.27$	$P = 0.003$*
One-half of questions (difficulty index < 0.872, $n = 69$ questions)			
3-option	0.74 (0.13)	0.36 (0.11)	0.36 (0.13)
4-option	0.72 (0.13)	0.35 (0.11)	0.36 (0.13)
5-option	0.69 (0.13)	0.37 (0.11)	0.38 (0.15)
Test: means are same	$P < 0.0001$*	$P = 0.47$	$P = 0.50$

Values of indexes are means (1 SD). Each null hypothesis, that the means of the indexes were not affected by the number of options provided per question, was tested using a mixed-model analysis that also included question identity and student group as random effects. In all cases, question identity was significant at $P < 0.0005$. The complete data set was analyzed (*top*) in addition to separate analyses of the least difficult (*middle*) and most difficult questions (*bottom*). *Significant difference.

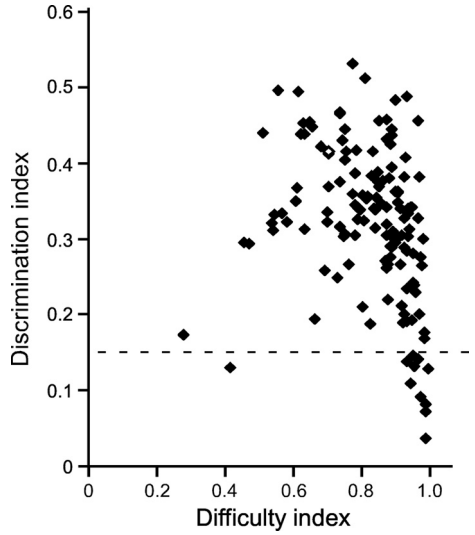

Fig. 2. The discrimination index (point biserial) for each question (averaged for the 3 versions with different numbers of options) is plotted against the difficulty index for the same question (also averaged for the 3 versions) (n = 138 questions). Most of the questions with an average discrimination index below 0.15 (indicated by the dashed line) had a high difficulty index (close to 1.0). The question used in Fig. 3 is indicated by the symbol with a white center.

correctly than the students who had the same question with three options (64% compared with 78% for this specific question), there was no corresponding monotonic increase in either discrimination index: the point biserial discrimination index decreased from the three-option to the five-option version (0.40 compared with 0.45), and the upper-lower discrimination index was highest for the four-option version (Fig. 3*A*). When the response pattern is further broken down by examining the option choices between students who are divided into four groups on the basis of their total score on that exam (quartiles), the pattern is consistent across different numbers of options (Fig. 3*B*). In all cases, the students who scored higher on the exam (4th quartile) were much more likely to select the correct option and less likely to select the distractors than students in the other three quartiles (Fig. 3*B*). Furthermore, the correct option showed a monotonic increase (in the proportion of students selecting that option) with quartile in all cases, whereas the distractors (incorrect options) tended to decrease with quartile in all cases (Fig. 3*B*).

To evaluate whether these patterns in the discrimination indexes and difficulty index were driven by some of the extremes in the data, the set of questions was divided in half (by the magnitude of the difficulty index), and the mixed-model analyses were performed for each half of the data. For the easier half of the questions (questions for which $>$ 87.2% of the students answered correctly, averaging over the 3 versions), the results were qualitatively the same as seen in the full data set of questions (Table 1). For the more difficult half of the questions (questions for which $<$87.2% of the students answered correctly, averaging over the 3 versions), the difficulty index, but neither discrimination index, showed a statistically significant effect of the number of options (Table 1).

The KR20 score was calculated for each of the 18 versions of the exams (3 exams $\times$ 2 sections $\times$ 3 versions for each exam). The average KR20 was 0.70, with a range from 0.61 to 0.81 (n = 18). The KR20 is an indicator of the reliability of an exam as a whole. KR20 is expected to be ~0.675 for a 30-question exam with 4 choices/question [calculated following Ebel (9)]. Therefore, these exams may be considered reliable within accepted standards.

Presumably the effects of eliminating distractors would depend on how frequently they were chosen by students when available, and so we used the selection pattern of the four distractors in the five-option case to estimate the preference for the eliminated vs. noneliminated distractors for the three-option case (*Eq. 4*). Not surprisingly, for those questions in which the eliminated options were less frequently selected when available (metric $>$ 0), eliminating them had little effect on either the difficulty index or discrimination index (Fig. 4, right-hand sides of graphs), when comparing between the three-option case and the five-option case for any question. When the eliminated options were preferred when available in the five-option version (metric $<$ 0), a larger number of students got the three-option version of the question correct than those with the five-option version of the question (Fig. 4*A*) (t-test comparing difficulty index difference for metric $<$ 0 and metric $>$ 0 using Satterthwaite method for unequal variances: P = 0.02). For these same questions, there was no statistically significant difference in the point biserial discrimination index response whether the distractors were preferred or not (Fig. 4*C*) (t-test comparing point biserial index difference for metric $<$ 0 and metric $>$ 0: P = 0.1). This result is not consistent with a hypothesis that eliminated distractors were preferred by the weaker students rather than the stronger students in the class (as indicated by their total score on that exam) (Fig. 4*D*). Results for the upper-lower discrimination index differences (not shown in Fig. 4) were qualitatively similar to those for the point biserial index differences (t-test comparing upper-lower index difference for metric $<$ 0 and metric $>$ 0: P = 0.2).

For each of the exams during the quarter, some of the questions were identical between the two class sections: for *exam 1*, 15 questions were identical, for *exam 2*, 13 questions were identical, and for *exam 3*, 1 question was identical. Therefore, during the quarter as a whole, 29 questions (out of the 138) were identical between the two sections of the class. The difficulty index for the 29 questions when compared between the two sections was highly correlated (r = 0.89–0.95 comparing the 5-option with 5-option, 4-option with 4-option, 3-option with 3-option). This means that approximately the same proportion of the students answered that same question correctly even though it was embedded in a slightly different group of questions. In contrast, the point biserial discrimination index was not highly correlated for the same 29 questions when compared between the two sections (r = 0.04–0.42 comparing the 5-option with 5-option, 4-option with 4-option, 3-option with 3-option). These trends mirror the correlations observed between the three-, four-, and five-option versions within exams within sections (Fig. 5) and suggest that, in general, the discrimination indexes of a question are less reproducible than the difficulty index.

Twenty-seven of the 138 questions were categorized as quantitative in nature, requiring the student to perform a calculation or compare magnitudes to select the correct answer. The quantitative questions were more likely to have a higher discrimination index for the five-option version compared with the three-option version of the same question than the non-quantitative questions: 78% (21/27) of the quantitative ques-

A

Question text: One difference between the pulmonary and systemic circulation is that the pulmonary circulation		3 options	4 options	5 options
option 1:	**has a lower pressure than the systemic circulation. (correct answer)**	78%	68%	64%
option 2:	has a lower volume flow rate than the systemic circulation.	16%	16%	13%
option 3:	has a larger volume than the systemic circulation.	6%	2%	4%
option 4:	does not have capillary beds while the systemic circulation does.	-	14%	15%
option 5:	has portal systems while the systemic circulation does not.	-	-	4%
	point biserial	0.45	0.39	0.40
	upper-lower	0.48	0.54	0.51
	n	110	108	109

B

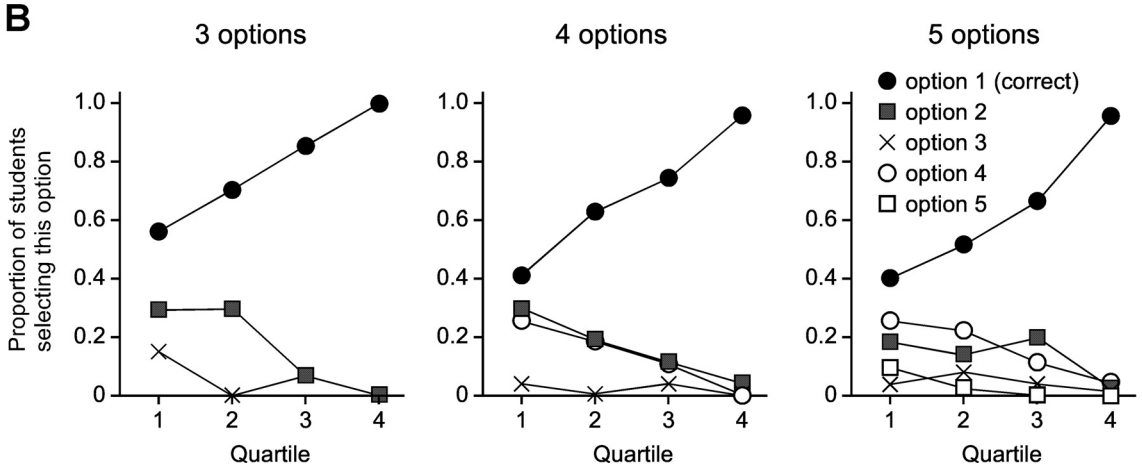

Fig. 3. *A*: a representative question, arbitrarily chosen, is provided, showing the five possible options (*option 1* is the correct answer). The percentage of the students who picked each option is provided on the *right*, for the three different versions of that question (with different numbers of options). The point biserial, upper-lower discrimination index, and *n* number of students for each option version of this question are provided. *B*: the data from *A* are further broken down by student quartile; *quartile 4* is the top 25% of the students (based on their overall score on that exam). For each quartile (~27 students), the proportion of students in that quartile picking each option is displayed.

tions had a higher discrimination index for the five-option version, whereas 56% (62/111) of the nonquantitative questions had a higher discrimination index for the five-option version. These proportions are significantly different ($P = 0.0482$, Fisher's exact test, two-sided). Although statistically significant, the overall increase in discrimination index was very small: for the quantitative questions, the upper-lower discrimination index increased by an average of 0.06 in the five-option case compared with the three-option case ($N = 27$), whereas, in the nonquantitative questions, the upper-lower discrimination index increased by an average of 0.02 ($N = 111$).

The overall length of a question (the stem and options) was evaluated by counting the number of words. The total number of words in a question ranged from 11 to 189, with an average of 61 words ($N = 138$). The upper-lower discrimination index was significantly higher for a question with more words ($P = 0.0002$), but not the number of options ($P = 0.8$). In contrast, the point biserial discrimination index was not significantly affected by either the total number of words in a question ($P = 0.09$) or the number of options ($P = 0.4$). A longer question did tend to be more difficult ($P < 0.0001$), but was not significantly affected by the number of options in addition ($P = 0.1$).

The total number of words in a question are divided between the stem and the options. The proportion of the words in the options ranged from 5 to 97% of the total for a question, with an average of 53% ($N = 414$, considering all three versions for each of 138 questions). Unlike the total number of words, the division of words between the stem and options did not have a statistically significant effect on either of the discrimination indexes or the difficulty index, when evaluated in an analysis of covariance with the number of options ($P > 0.05$ in all cases).

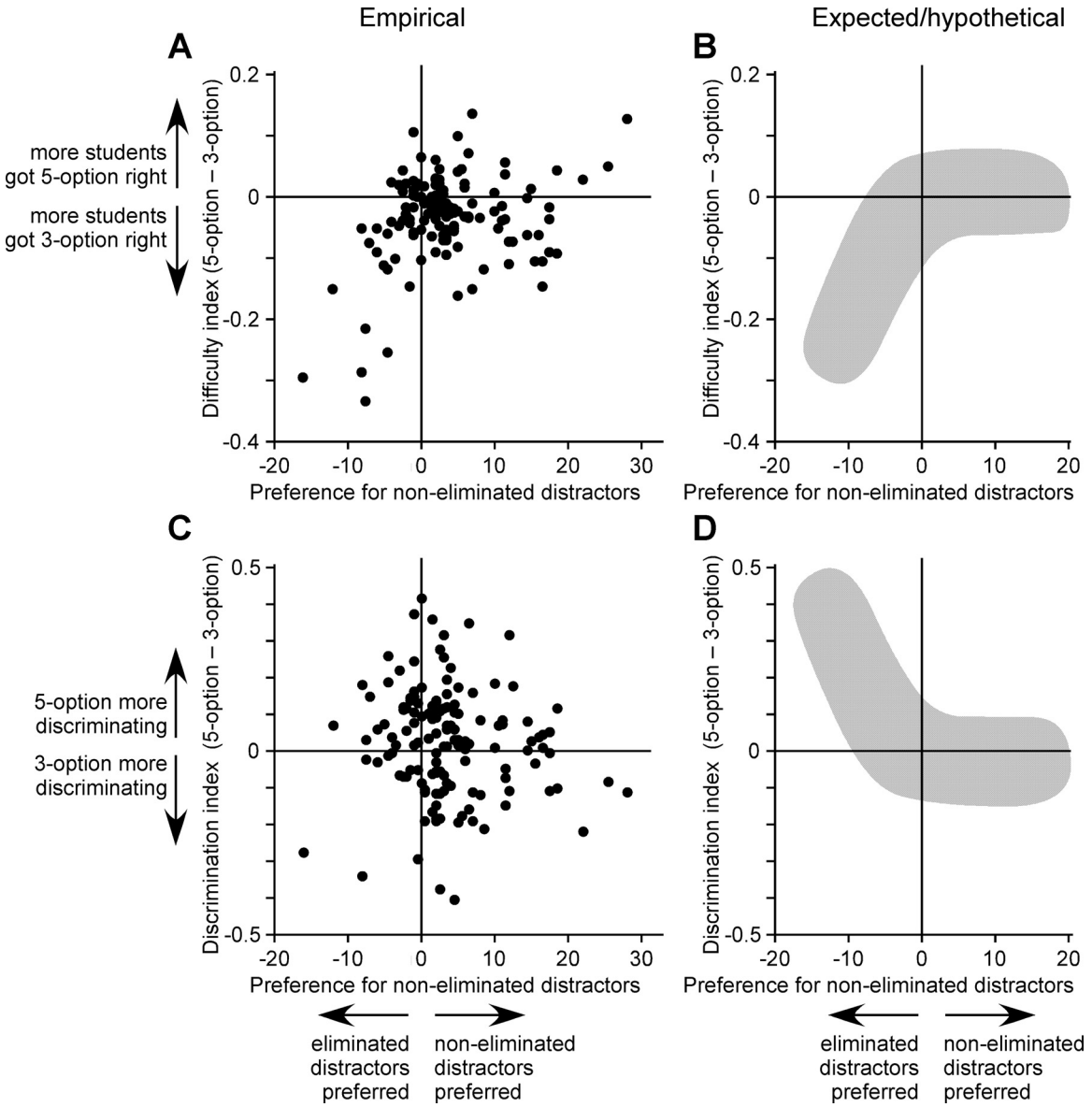

Fig. 4. *A*: the difficulty index was affected by the preference for the noneliminated distractors (the number of students who chose either of the noneliminated distractors minus the number of students who chose either of the eliminated distractors, all in the 5-option version). Each point is for 1 of the 138 questions. *B*: a hypothetical curve showing the approximate expectation for the relationship shown in *A*. The stronger the preference for the noneliminated options, the less likely it is that the difficulty index would change between the five- and three-option versions of each question. However, for those questions for which there is a preference for the eliminated options (a negative value on the *x*-axis), the expectation would be that more students would get the three-option version correct than the five-option version. *C*: the discrimination index (point biserial) did not show a significant trend with preference for the noneliminated distractors. Each point is for 1 of the 138 questions. *D*: a hypothetical curve showing the approximate expectation for the relationship shown in *C*: the stronger the preference for the noneliminated options, the less likely it is that the discrimination index would change between the five- and three-option versions of each question. However, if the eliminated options were preferred by the weaker students, then the discrimination index would be higher for the five-option version of the question.

DISCUSSION

Our results for testing in a large undergraduate biology class in human physiology confirm some of the earlier reports from the education literature that having more than three options for a multiple-choice question may not improve summative assessment. Specifically, the point biserial discrimination index did not show a statistically significant improvement when more than three options were provided, and while a statistically significant increase in the upper-lower discrimination index was detected, this was only a trivial gain, and then only for the easier questions. The experimental design used in this study, in which the same questions with different numbers of options were given to different subsets of the class at the same time as part of regular testing in the classroom, provides a more rigorous evaluation of the number of options than is possible when comparing between years, when there can be other confounding variables. Our results suggest that, in a realistic setting, with highly motivated and senior undergraduate students, three options were sufficient for the types of questions and options used in this study.

More students selected the correct answer when there were fewer distractors, regardless of the difficulty of the question, or

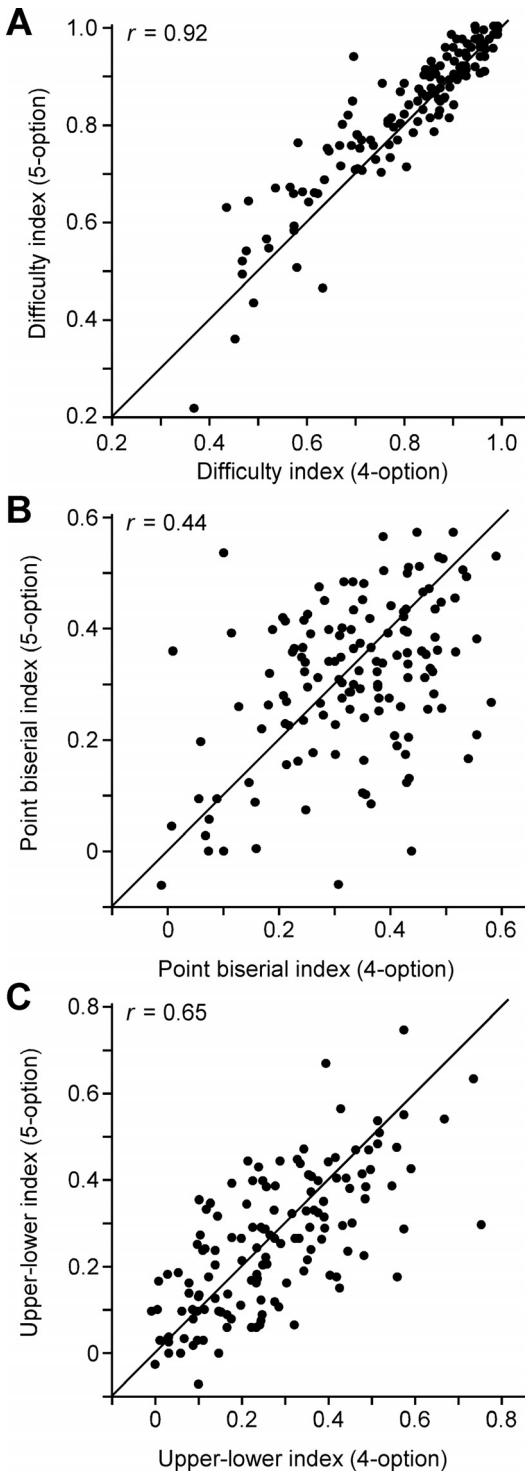

Fig. 5. *A*: the difficulty index was highly correlated ($r = 0.92$) between the four-option and five-option versions of the same questions ($n = 138$ questions). Each point is for a single question. *B*: the point biserial discrimination index was only weakly correlated ($r = 0.44$) between the four-option and five-option versions ($n = 138$ questions). *C*: the upper-lower discrimination index had a greater correlation ($r = 0.65$) between the four-option and five-option versions than the point biserial discrimination index ($n = 138$ questions). Very similar correlations were observed between three- and four-option versions (not shown).

the magnitude of the discrimination index. This observed pattern suggests that guessing does place a small role in selecting the answer. The small but significant difference in the proportion of students who answered a question correctly with different numbers of options may be used to make a rough estimate of how many students are guessing. Consider the very simple case in which the students can be divided into two groups: those who know the answer to a question (and will answer it correctly), and those who do not know the answer (and are equally likely to pick any of the options). The proportion of students who know the answer, *K*, and the proportion of students who do not know the answer, *D*, add up to 100%. The proportion of students who answer a question correctly, *C*, will be

$$C = K + \frac{D}{\text{number of options}} \quad (5)$$

For example, if 50% of the students know the answer, and the other 50% are guessing randomly between five options, the expectation would be that 50% + (50%/5) or 60% would answer the question correctly, whereas, if there are only four options, the expectation would be that 50% + (50%/4) or 62% would answer the question correctly. The expected outcome for this simple situation is shown in Fig. 6. The best fit for the observed data (difficulty indexes in Table 1) is found for a value of 76% for the percentage of students who know the answer (the best fit identified by minimizing the sum of squares between the observed and theoretical values). If 76% of the students know the answer (and answer the question correctly) and the remaining 24% of the students do not know the answer (and randomly select one of the options), the expected proportions are 84, 82, and 81% for three, four, and five options, respectively (compared with the empirically observed proportion of 84, 82, and 80%). This suggests that, if this set of simple assumptions is approximately valid, then, on average, ~24% of the students are guessing for any question on these exams.

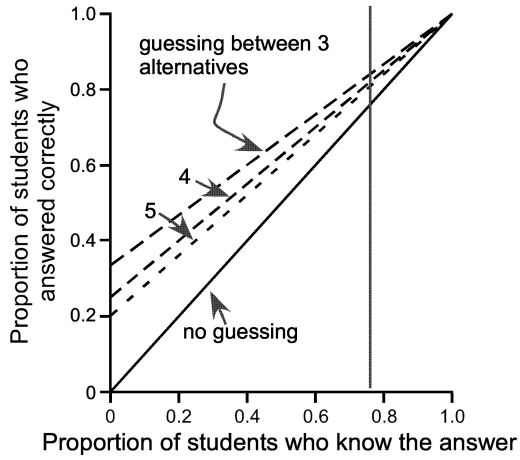

Fig. 6. Predictions of the proportions of students who answer a question correctly, assuming that a fixed proportion of students know the answer and answer the question correctly (*x*-axis), whereas the remaining students either *1*) randomly guess between three alternatives (so one-third of these remaining students will guess the correct answer); *2*) randomly guess between four alternatives; *3*) randomly guess between five alternatives; or *4*) do not guess at all. The vertical line indicates the case in which 76% of the students know the answer.

Note that the overall number of questions could be increased to counteract this increased guessing of correct answers with fewer alternatives. For the specific case of 30 questions with 5 options per question, increasing the number of questions to 44 while decreasing the number of options per question to 3 would offset the guessing (for 30 questions with 5 options there are 5×30 different response patterns; setting this equal to $3 \times X$ different response patterns expected from X questions with 3 options, and solving for X results in $X = 44$). The ability of an exam to protect against guessing is sometimes referred to as the "power" of an exam (11, 29). Note that, in this specific example, there is a smaller total number of options in the three-option case ($44 \times 3 = 132$) than in the five-option case ($30 \times 5 = 150$), even though they have the same power. If the time needed for an exam is directly proportional to the total number of options, then the three-option case in this specific example will take less time for the same power.

The similarity of the difficulty index for questions replicated between sections is interesting, because it suggests that a question had a level of difficulty that is somewhat independent of the rest of the exam. This may have been helped by the standard practice that, for all exams, care was taken so that the text provided in a question did not provide clues to the answers in other questions within the same exam. In contrast, it was surprising how dissimilar the discrimination indexes were for the same questions when compared between sections. It is not caused by the different background supplied by the rest of the questions, because the discrimination indexes were also only weakly correlated between the three-, four-, and five-option versions in the same set of questions within a single section (Fig. 5). There was no obvious pattern in which questions tended to have a higher correlation. Therefore, we conclude that the magnitude of a discrimination index for a given question is less reproducible than the magnitude of the difficulty index (at least for these questions in these exams). While only two types of discrimination index were used (point biserial correlation coefficient and upper-lower), the alternative discrimination indexes [e.g., Flanagan's coefficient, Davis' coefficient (10)] are mathematically related to each other, and there is no reason to expect different behavior from these other discrimination indexes. Therefore, while discrimination indexes are a useful metric for instructors to check, our results suggest that small differences in magnitude are not cause for concern. For example, in the present study, only 62% (18/29) of the point biserial discrimination indexes from the same questions given to the different sections were within 0.1 of each other in magnitude (comparing 5-option versions of the questions).

If a distractor is rarely picked, then its elimination from a set of options would not be expected to influence either the difficulty or a discrimination index (e.g., Fig. 4, *B* and *D*, *right*). Rather than simply characterize a distractor as being selected or not, we used a metric to describe how frequently the noneliminated distractors were selected (subtracting the number of students who chose either of the eliminated distractors from the number of students who chose either of the noneliminated distractors, all in the 5-option version). Similar to our expectations, we did find that elimination of less frequently picked distractors did not influence either the difficulty or the discrimination index (e.g., Fig. 4, *A* and *C*, *right*). We also found that elimination of more frequently picked distractors tended to increase the number of students who answered a question correctly (Fig. 4, *A* and *B*, *left*). Although we did not observe a statistically significant effect on a discrimination index when preferred distractors were eliminated, this pattern (Fig. 4*C*, *left*) suggests that, if in general the preferred distractors were usually the ones eliminated, we might have observed a significant effect. The low reproducibility of the discrimination indexes (e.g., Fig. 5) would make it more difficult to observe the expected pattern (Fig. 4*C*).

Our results confirm the reports in the literature that some options perform as weak distractors. A weak or nonfunctioning distractor is often defined as one that is chosen by fewer than 5% of the examinees (4, 14). While all of the distractors used in the questions in this study were considered to have merit (corresponded to a misconception or misunderstanding that some students have had about the material in our experience), 31% of the distractors in the three-option versions of the questions were chosen by fewer than 5% of the examinees, and 73% of the distractors in the five-option versions of the questions. Haladyna and Downing (14) suggest that three options per item may be a natural limit for writers, who just cannot think of any additional good distractors, but another interpretation is simply that some misconceptions are not widely held, despite being voiced by some students.

Obviously there might be specific times when a larger number of options is warranted. For example, an instructor might be interested in evaluating more than two common misconceptions associated with a specific point addressed by the question. The instructor may identify the prevalence of these misconceptions from the frequency of different options chosen by the students. Suggestions for how multiple-choice questions may be developed as diagnostic tools to expose students' misconceptions are described by Treagust (26).

In addition, in a theoretical treatment, Grier (11) described how the optimal number of options per question will increase as a greater proportion of the time is spent reading the stem of the questions (vs. the options). The optimum number of options increases the more time that it takes to read the stems of the questions; for example, if it takes twice as long to read the stem as one of the possible answers, the optimum number of options increases to 4, and if it takes 10 times as long to read the stem as one of the possible answers, the optimum number of options is 9 (11). While Grier's analysis refers to a testwise strategy in which the total amount of time for the exam is considered fixed, it provides an intriguing comparison to the results of Case et al. (3), which are otherwise a bit anomalous, in which they found that a large number of options (9–23 options per question) resulted in a higher discrimination index (0.22 vs. 0.18) compared with the five-option versions of the same questions. In this latter example (3), the bulk of the reading was in the questions, in which the patient's symptoms and circumstances were explained, and the options were short alternative diagnoses.

Motivated by these reports (3, 11), in which a larger number of options led to a more discriminating exam or more discriminating questions when the stem of a question was long or time-consuming to read relative to the options, we tested this prediction using our exam questions. We made the simplifying assumption that the time spent on part of a question could be approximated by the number of words of that part. While a larger total number of words in a question (stem + options)

was associated with a slightly more difficult and discriminating question (the upper-lower discrimination index, but not the point biserial), there was no effect seen on the difficulty or discrimination of a question based on the distribution of the words between the stem and the options.

Multiple-choice questions differ in a number of additional ways, such as subject matter, Bloom's level, or the types of skills on which students need to draw to select the correct answer. These different types of multiple-choice questions might result in a different expectation for the optimal number of options per question. We considered whether multiple-choice questions that were quantitative in nature, such that students needed to make calculations to answer the question, would show a different optimal number of options than nonquantitative questions (Fig. 3 shows an example of a nonquantitative question). A significantly larger proportion of the quantitative questions showed an increase in the magnitude of the upper-lower discrimination index with more options, compared with the nonquantitative questions. This may indicate that it is easier to predict common misconceptions (such as problems in quantitative reasoning) for quantitative questions, and explicitly address these common misconceptions as distractors. However, because this significant effect was small in magnitude, this small gain in discrimination does not by itself make a strong case for including more options to improve assessment. It remains to be seen if other types of multiple-choice questions may be identified that are more effective with larger numbers of options.

Ebel (9) pointed out that a multiple-choice exam with fewer options per question will have a lower reliability, if the number of questions is kept fixed, and he provided a simple formula for calculating the expected reliability. Using his formula, for an exam with 30 questions and an average of 4 options per question, the reliability is theoretically predicted to be 68%. We had a reliability (KR20) of 0.70 for our exams (with an average of 4 options per question). A value of 0.70 is often used as a target for reliability, although this is also a function of exam length and other factors (5). In another theoretical analysis, Grier (12) expanded on Tversky's theoretical treatment (29) to show that the reliability of an exam (the KR21) is maximized for three options, if the total number of options in the exam is held constant.

In conclusion, the results of this study provide good news for both instructors and students. For an instructor, having fewer options per question (3 rather than 4 or 5) makes writing exam questions less laborious. For students, having fewer options requires less exam time reading and evaluating for each question. After decreasing the number of options, an instructor has a choice to include more questions in the same amount of testing time, which will increase exam reliability, or to have less rushed students (20). It is worth pointing out that there is general acknowledgment that the quality of the distractors is more important than the number of distractors (15). The ultimate objective here is not to simply have fewer distractors, but to have fewer but well-functioning distractors. In practice, faculty may find it helpful to start with more distractors and then, following item analysis, figure out which or why some distractors should be deleted or amended to improve the assessment. Instructors should feel emboldened, rather than compelled, to consider providing fewer options. Instructors may choose to include more options if they wish to decrease the impact of guessing, or to evaluate the frequency of certain misconceptions.

Glossary

Difficulty index	Percentage of students who answer a question correctly
Discrimination index	How likely the higher-scoring students are to answer a particular question correctly
Distractor	Incorrect option
Kuder-Richardson 20 (KR20)	Estimate of reliability, *Eq. 3*
Kuder-Richardson 21 (KR21)	Estimate of reliability with simpler formula
Point biserial correlation coefficient	One type of discrimination index, *Eq. 1*
Upper-lower	Another type of discrimination index, *Eq. 2*

ACKNOWLEDGMENTS

A talk associated with this work was presented at the national meeting of the Society of Integrative and Comparative Biology in 2015 in West Palm Beach, FL. Statistical advice from Dr. Megan T. Smith, University of California-Irvine (UCI) Center for Statistical Consulting, is gratefully acknowledged. Constructive suggestions from Dr. Kameryn Denaro (UCI Division of Teaching Excellence and Innovation) improved the paper.

GRANTS

Partial funding for A. Macias-Muñoz was provided by National Science Foundation Bio/computational Evolution in Action Consortium, Division of Biological Infrastructure no. 0939454.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the authors.

AUTHOR CONTRIBUTIONS

C.L. conceived and designed research; C.L. and A.M.-M. performed experiments; C.L. analyzed data; C.L. and A.M.-M. interpreted results of experiments; C.L. prepared figures; C.L. drafted manuscript; C.L. and A.M.-M. edited and revised manuscript; C.L. and A.M.-M. approved final version of manuscript.

REFERENCES

1. **Brennan RL.** A generalized upper-lower item discrimination index. *Educ Psychol Meas* 32: 289–303, 1972. doi:10.1177/001316447203200206.
2. **Case SM, Swanson DB.** *Constructing Written Test Questions for the Basic and Clinical Sciences* (3rd ed.). Philadelphia, PA: National Board of Medical Examiners, 2002.
3. **Case SM, Swanson DB, Ripkey DR.** Comparison of items in five-option and extended-matching formats for assessment of diagnostic skills. *Acad Med* 69, Suppl: S1–S3, 1994. doi:10.1097/00001888-199410000-00023.
4. **Cizek GJ, O'Day DM.** Further investigation of nonfunctioning options in multiple-choice test items. *Educ Psychol Meas* 54: 861–872, 1994. doi:10.1177/0013164494054004002.
5. **Cortina JM.** What is coefficient alpha—an examination of theory and applications. *J Appl Psychol* 78: 98–104, 1993. doi:10.1037/0021-9010.78.1.98.
6. **Costin F.** The optimal number of alternatives in multiple-choice achievement tests: some empirical evidence for a mathematical proof. *Educ Psychol Meas* 30: 353–358, 1970. doi:10.1177/001316447003000217.
7. **Costin F.** Three-choice versus four-choice items: implications for reliability and validity of objective achievement tests. *Educ Psychol Meas* 32: 1035–1038, 1972. doi:10.1177/001316447203200419.
8. **Dirks C, Wenderoth MP, Withers M.** *Assessment in the College Science Classroom.* New York: Freeman, 2014.

9. **Ebel RL.** Expected reliability as a function of choices per item. *Educ Psychol Meas* 29: 565–570, 1969. doi:10.1177/001316446902900302.
10. **Ebel RL, Frisbie DA.** *Essentials of Educational Measurement* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall, Inc, 1986.
11. **Grier B.** Optimal number of alternatives at a choice point with travel time considered. *J Math Psychol* 14: 91–97, 1976. doi:10.1016/0022-2496(76)90016-X.
12. **Grier JB.** Number of alternatives for optimum test reliability. *J Educ Meas* 12: 109–112, 1975. doi:10.1111/j.1745-3984.1975.tb01013.x.
13. **Haladyna TM.** *Developing and Validating Multiple-choice Test Items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum, 2004.
14. **Haladyna TM, Downing SM.** How many options is enough for a multiple-choice test item? *Educ Psychol Meas* 53: 999–1010, 1993. doi:10.1177/0013164493053004013.
15. **Haladyna TM, Downing SM.** Validity of a taxonomy of multiple-choice item-writing rules. *Appl Meas Educ* 2: 51–78, 1989. doi:10.1207/s15324818ame0201_4.
16. **Landrum RE, Cashin JR, Theis KS.** More evidence in favor of three-option multiple-choice tests. *Educ Psychol Meas* 53: 771–778, 1993. doi:10.1177/0013164493053003021.
17. **Levine MV, Drasgow F.** The relation between incorrect option choice and estimated ability. *Educ Psychol Meas* 43: 675–685, 1983. doi:10.1177/001316448304300301.
18. **Lord FM.** Optimal number of choices per item: a comparison of four approaches. *J Educ Meas* 14: 33–38, 1977. doi:10.1111/j.1745-3984.1977.tb00026.x.
19. **Loudon C, Macias-Muñoz A.** Multiple-choice testing in physiology: are we providing too many alternative answers per question? *Integr Comp Biol* 55: E115, 2015.
20. **Owen SV, Froman RD.** What's wrong with three-option multiple choice items? *Educ Psychol Meas* 47: 513–522, 1987. doi:10.1177/0013164487472027.
21. **Rodriguez MC.** Three options are optimal for multiple-choice items: a meta-analysis of 80 years of research. *Educ Meas* 24: 3–13, 2005. doi:10.1111/j.1745-3992.2005.00006.x.
22. **Rogers WT, Harley D.** An empirical comparison of three-and four-choice items and tests: susceptibility to testwiseness and internal consistency reliability. *Educ Psychol Meas* 59: 234–247, 1999. doi:10.1177/00131649921969820.
23. **Sidick JT, Barrett GV, Doverspike D.** Three-alternative multiple choice tests: an attractive option. *Person Psychol* 47: 829–835, 1994. doi:10.1111/j.1744-6570.1994.tb01579.x.
24. **Tarrant M, Ware J.** A comparison of the psychometric properties of three- and four-option multiple-choice questions in nursing assessments. *Nurse Educ Today* 30: 539–543, 2010. doi:10.1016/j.nedt.2009.11.002.
25. **Tarrant M, Ware J.** A framework for improving the quality of multiple-choice assessments. *Nurse Educ* 37: 98–104, 2012. doi:10.1097/NNE.0b013e31825041d0.
26. **Treagust DF.** Development and use of diagnostic tests to evaluate students' misconceptions in science. *Int J Sci Educ* 10: 159–169, 1988. doi:10.1080/0950069880100204.
27. **Trevisan MS, Sax G, Michael WB.** The effects of the number of options per item and student ability on test validity and reliability. *Educ Psychol Meas* 51: 829–837, 1991. doi:10.1177/001316449105100404.
28. **Trevisan MS, Sax G, Michael WB.** Estimating the optimum number of options per item using an incremental option paradigm. *Educ Psychol Meas* 54: 86–91, 1994. doi:10.1177/0013164494054001008.
29. **Tversky A.** On the optimal number of alternatives at a choice point. *J Math Psychol* 1: 386–391, 1964. doi:10.1016/0022-2496(64)90010-0.